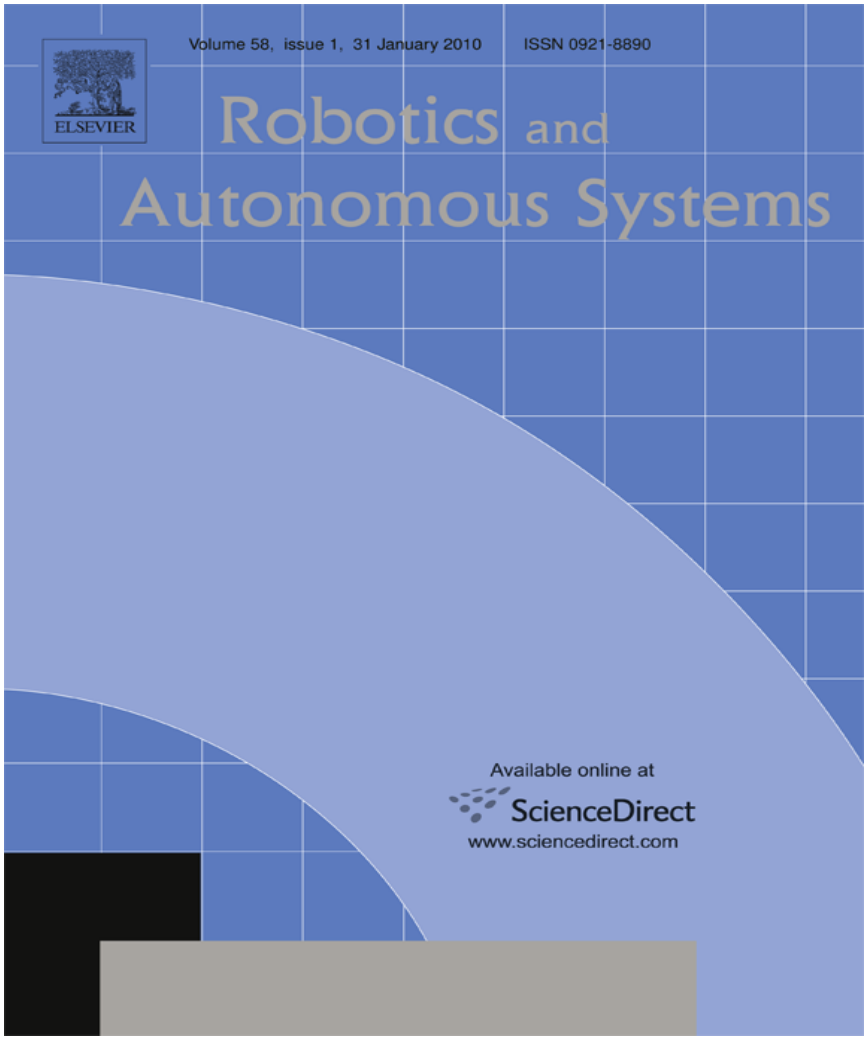


Robotics and Autonomous Systems 58 (2010) 81–96

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

A realistic benchmark for visual indoor place recognition☆

A. Pronobis [a,*], B. Caputo [b], P. Jensfelt [a], H.I. Christensen [c]

[a] Centre for Autonomous Systems, The Royal Institute of Technology, SE-100 44 Stockholm, Sweden
[b] IDIAP Research Institute, 1920 Martigny, Switzerland EPFL, 1015 Lausanne, Switzerland
[c] College of Computing, Georgia Institute of Technology, Atlanta, GA 30332-0760, USA

ARTICLE INFO

Article history:
Received 2 September 2008
Received in revised form
21 July 2009
Accepted 30 July 2009
Available online 4 August 2009

Keywords:
Visual place recognition
Robot topological localization
Standard robotic benchmark

ABSTRACT

An important competence for a mobile robot system is the ability to localize and perform context interpretation. This is required to perform basic navigation and to facilitate local specific services. Recent advances in vision have made this modality a viable alternative to the traditional range sensors, and visual place recognition algorithms emerged as a useful and widely applied tool for obtaining information about robot's position. Several place recognition methods have been proposed using vision alone or combined with sonar and/or laser. This research calls for standard benchmark datasets for development, evaluation and comparison of solutions. To this end, this paper presents two carefully designed and annotated image databases augmented with an experimental procedure and extensive baseline evaluation. The databases were gathered in an uncontrolled indoor office environment using two mobile robots and a standard camera. The acquisition spanned across a time range of several months and different illumination and weather conditions. Thus, the databases are very well suited for evaluating the robustness of algorithms with respect to a broad range of variations, often occurring in real-world settings. We thoroughly assessed the databases with a purely appearance-based place recognition method based on support vector machines and two types of rich visual features (global and local).

1. Introduction

A fundamental competence for an autonomous agent is to know its position in the world. Providing mobile robots with abilities to build an internal representation of space and obtain robust information about their location therein can be considered as one of the most urgent problems. The topic is vastly researched. This resulted, over the years, in a broad range of approaches spanning from purely metric [1–3] to topological [4–6] and hybrid [7,8]. As robots break down the fences and start to interact with people [9] and operate in large-scale environments [6,5], topological models are gaining popularity for augmenting or replacing purely metric space representations. In particular, the research on topological mapping has pushed methods for place recognition. Scalability, loop closing and the kidnapped robot problem have been at the forefront of the issues to be addressed.

Traditionally, sonar and/or laser have been the sensory modalities of choice for place recognition and topological localization [10,11]. The assumption that the world can be represented in terms of two-dimensional geometrical information allowed for many practical implementations. Yet, the inability to capture many aspects of complex realistic environments leads to the problem of perceptual aliasing [12] and greatly limits the usefulness of purely geometrical methods. Recent advances in vision have made this modality emerge as a natural and viable solution. Vision provides richer sensory input allowing for better discrimination. It opens new possibilities for building cognitive systems, actively relying on the semantic context. Not unimportant is the cost effectiveness, portability and popularity of visual sensors. As a result, this research line is attracting more and more attention, and several methods have been proposed using vision alone [13–15,6], or combined with more traditional range sensors [16–18].

In spite of large progress, vision-based localization still represents a major challenge. First of all, visual information tends to be noisy and difficult to interpret. The visual appearance of places varies in time because of illumination changes (day and night, artificial light on and off) and because of human activities (furniture moved around, objects being taken out of drawers, and so on). Thus, the solutions must be highly robust, provide good generalization abilities and in general be adaptive. Additionally, the application puts strong constraints on the computational

☆ A preliminary version of the experimental evaluation reported in this work was presented in: A. Pronobis, B. Caputo, P. Jensfelt, and H.I. Christensen. A discriminative approach to robust visual place recognition, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS'06, Beijing, China, October 2006.

* Corresponding author.
E-mail addresses: pronobis@csc.kth.se (A. Pronobis), bcaputo@idiap.ch (B. Caputo), patric@csc.kth.se (P. Jensfelt), hic@cc.gatech.edu (H.I. Christensen).

0921-8890/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.robot.2009.07.025

complexity, and the increased resolution and dimensionality of the visual data still constitute a problem.

The fact that so many different parameters influence the accuracy of a vision-based localization system is another challenge itself, especially burdensome at the design stage. As the results depend greatly on the choice of training and test input data, which are unstable over time, it is hard to measure the influence of the different parameters on the overall performance of the system. For the same reason, it becomes nearly impossible to compare fairly solutions which are usually evaluated in different environments, under different conditions and with different assumptions. This is a major obstacle slowing down progress in the field. There is a need for standardized benchmarks and databases which would allow for fair comparisons, simplify the experimental process and boost development of new solutions.

Databases are heavily exploited in the computer vision community, especially for object recognition and categorization [19–21]. As the community acknowledges the need for benchmarking, a lot of attention is directed towards designing new datasets, reflecting the increasing capabilities of visual algorithms [22]. Also in robotics, research on simultaneous localization and mapping (SLAM) makes use of several publicly available datasets [23,24]. Still, no database emerged as a standard benchmark for visual place recognition applied to robot localization.

This paper aims at filling this gap and presents a benchmark consisting of two different image databases gathered in the same indoor environment. The databases are augmented with an experimental procedure as well as extensive baseline evaluation. The datasets were carefully designed and later annotated. Three different imaging devices were used for acquisition (two mobile robot platforms and a standard camera), resulting in data of different characteristics and quality. In order to create a realistic and challenging test bed, the acquisition process was performed in an uncontrolled typical office environment, under various illumination and weather conditions (sunny, cloudy, night) and over a significant span of time. All of these make the databases very well suited for evaluating robustness of visual place recognition algorithms, applied to the problem of robot topological localization, in the presence of different types of variations often occurring in real-world indoor settings.

An important component when providing the community with a new collection of data is to provide a baseline evaluation that illustrates the nature of the dataset (see Section 5.1 for explanation). We thoroughly assessed the databases with a purely appearance-based place recognition method. The method uses two types of image descriptors, local and global, in order to extract rich visual information. Both descriptors have shown remarkable performances, coupled with computational efficiency on challenging object recognition scenarios [25,26]. The classification step is performed using support vector machines (SVMs) [27] and specialized kernels are used for each descriptor. Results show that the method is able to recognize places with high precision and robustness under varying illumination conditions, even when training on images from one camera device and testing on another.

The rest of the paper is organized as follows: after a review of the related literature (Section 2), we discuss the problem and challenges we addressed with the benchmark (Section 3). Then, Section 4 gives a detailed description of the data acquisition process and scenario and presents the acquisition results. Finally, the algorithm used for the baseline evaluation as well as the experimental procedure are described in Section 5, and the experimental results are given in Section 6. The paper concludes with a summary (Section 7).

2. Related work

Place recognition and topological localization are vastly researched topics in the robotic community,where vision and laser range sensors are usually the privileged modalities. Although laser-based solutions have proven to be successful for certain tasks [11], their limitations inspired many researchers to turn towards vision which nowadays becomes tractable in real-time applications. The available methods employ either perspective [13,28,29] or omnidirectional cameras [30,31,4,32–35]. The main differences between the approaches relate to the way the scene is perceived and thus the method used to extract characteristic features from the scene. Landmark-based techniques make use of either artificial or natural landmarks in order to extract information about a place. Mata et al. [36] proposed a system able to interpret information signs through its ability to read the text and recognize icons. Visually distinctive image regions were also used as landmarks [15]. Other solutions employed mainly local image features such as SIFT [25,33,14], SURF [37,34,35], also using the bag-of-words approach [29,38,6], or representation based on information extracted from local patches using Kernel PCA [28]. Global features are also commonly used for place recognition. Torralba et al. [39,13,40] suggested to use a representation called the "gist" of the scene, which is a vector of principal components of outputs of a bank of spatially organized filters applied to the image. Other approaches use color histograms [4,31], gradient orientation histograms [41], eigenspace representation of images [30], or Fourier coefficients of low frequency image components [32]. Recently, several authors observed that robustness and efficiency of the recognition system can be improved by combining information provided by both types of visual cues (global and local) [14,15,42]. Although vision-based localization methods are now commonly applied, it remains extremely difficult to compare the different approaches, as the evaluations presented by the authors usually follow different procedures and are performed on different sets of visual data.

There are a number of heavily used standard databases in robotics and computer vision. In robotics, these databases are used mainly for testing algorithms for SLAM [23,24] and mostly contain odometry and range sensor data. In the case of the computer vision community, the effort concentrated on creating standard benchmarks for such problems as object [19,20,22], action [43], scene [21], or texture recognition and categorization [44]. The MIT-CSAIL Database of Objects and Scenes [21] is a notable exception as it provides several image sequences acquired in both indoor and outdoor environments and was used to evaluate the performance of a visual place recognition system.

This paper makes an important contribution by providing annotated data from visual and laser range sensors together with an experimental procedure that can be followed in order to evaluate place recognition and localization systems. In contrast to the previously available benchmarking solutions, the databases contain several sets of images and image sequences acquired in the same environment under various conditions and over a significant span of time. This makes them perfect for evaluating robustness of the algorithms under dynamic variations that often occur in realistic settings. The introduction of standard benchmark databases has made an impact on the research on such problems as object categorization or SLAM, allowing different methods to be more fairly compared in the same scenario. The authors hope that the benchmark proposed in this paper will similarly influence the research on visual place recognition in the context of mobile robot localization.

3. Design strategy

This section defines and characterizes the problem that we address with the benchmark (Section 3.1) and analyzes the difficulties and open challenges in visual place recognition that have to be considered in a realistic scenario (Section 3.2).

3.1. Problem statement

Let us begin with a brief definition of a place and the place recognition problem that we will use throughout this paper. A place can be regarded as a usually nameable segment of a real-world environment distinguished due to different functionalities, appearances or artificial boundaries. In view of this definition, the place recognition or identification problem can be characterized as follows. Given a set of training sensory data, captured in each of the considered places, build models of the places reflecting their inherent properties. Next, when presented with new test data, unavailable during training, acquired in one of the same places, identify the place where the acquisition was performed (e.g. Barbara's office) based on the knowledge encoded in the models. This is different from the problem of place categorization where the task is to classify test data captured in a novel place as belonging to one of the place categories (e.g. an office). As the partition of space into different places can be based on several criteria, here we consider a supervised scenario where the algorithm has to distinguish between five areas of different functionalities, selected by a teacher.

This benchmark is designed to test the performance of a visual place recognition system on images acquired within an indoor office environment. As the primary scenario, we consider the case where a place recognition system is used to provide a mobile robot with information about its location. For this reason, part of the data presented in this paper was acquired using cameras mounted on mobile robot platforms. While designing the benchmark, we concentrated on testing the ability of a visual recognition system to identify a place based on one image only. This makes the problem harder, but also makes it possible to perform global localization where no prior knowledge about the position is available (e.g. in the case of the kidnapped robot problem). Spatial or temporal filtering can be used together with the presented methods to enhance performance.

We concentrate on indoor environments since in the considered scenario, they play a crucial role, being typical spaces for the interaction between humans and service robots or robotic assistants [9]. At the same time, office environments, just like home environments, constitute an important class of indoor spaces for robotic companions. In this benchmark, our aim is to provide datasets and experimental procedures that will allow for evaluating robustness of place recognition systems based on different types of visual cues to typical variations that occur in an indoor environment for the considered scenario. These include illumination changes, variations introduced by human activity and viewpoint changes. As a consequence, instead of providing datasets spanning over a very large portion of space, we provide image sequences acquired over a time span of several months, under various illumination conditions and using different devices. The proposed evaluation framework should allow for concluding that an algorithm robust to the variations captured in the benchmark data will be robust to similar types of variations within other indoor office environments.

The benchmark is designed for evaluating vision-based methods. We choose vision as sensory modality for several reasons. First, the visual sensor is very rich and, although also very noisy, provides great descriptive capabilities. This is crucial in indoor environments where other sensors, such as a laser range finder, suffer from the problem of perceptual aliasing (different places look the same [12]). Furthermore, the visual appearance of places encodes information about their semantics, which plays a major role in enabling systems to interact with the environment. Finally, in the era of cheap portable devices equipped with digital cameras, it is also one of the most affordable and commonly available solutions.

3.2. Challenges

Recognizing indoor places based on their visual appearance is a particularly challenging task. First of all, in the case of indoor environments, there is no obvious spatial layout that once observed could be used to distinguish between different places. Moreover, viewpoint variations cause the visual sensor to capture different aspects of the same place, which often can only be learned if enough training data are provided. At the same time, real-world environments are usually dynamic and their appearance changes over time. The visual recognition system must be robust to variations introduced by changing illumination as well as human activity. For a visual sensor, the same room might look different during the day, during sunny weather, under direct natural illumination, and at night with only artificial light turned on. Moreover, if the environment is being used, the fact that people appear in the images, objects are being moved or furniture relocated may greatly influence the performance of the system. All these issues were taken into consideration while designing this benchmark in order to create a realistic test bed.

4. Data acquisition

Based on the analysis of the problem presented in the previous section, we carefully designed and acquired two databases comprising images captured in the same indoor environment, but using different devices: the INDECS (INDoor Environment under Changing conditionS) database [45] and the IDOL (Image Database for rObot Localization) database [46]. This section describes the resulting data acquisition procedure. In the case of INDECS, we acquired images of the environment from a fixed set of points using a standard camera mounted on a tripod. The resolution of the images is high; this makes this database suitable for context-based object recognition. The IDOL database, instead, consists of image sequences recorded using two mobile robot platforms equipped with perspective cameras, and thus is well suited for experiments with robot localization. All three devices are shown in Fig. 1. The databases represent a different approach to the problem and can be used to analyze different properties of a place recognition system. The acquisition was performed under several different illumination settings and over a significant span of time. Both databases are publicly available and can be downloaded from http://www.csc.kth.se/~pronobis.

The rest of the section is organized as follows: Section 4.1 presents the acquisition scenario, as to say the environment where both databases were acquired. Then, Section 4.2 provides a description of the INDECS database, and Section 4.3 gives detailed information on the robot platforms and IDOL. Finally, we perform an analysis of the obtained data in Section 4.4.

4.1. Acquisition scenario

The acquisition was conducted within a five room subsection of a larger office environment of the Computer Vision and Active Perception Laboratory at the Royal Institute of Technology in Stockholm, Sweden. Each of the five rooms represents a different type of functional area: a one-person office, a two-person office, a kitchen, a corridor and a printer area (in fact a continuation of the corridor). The function that a room fulfills determines the furniture, objects and activity that is likely to be found there. Places like the corridor, the printer area and the kitchen can be regarded as public which implies that various people may be present. On the other hand, offices were imaged usually empty or with their owners at work. In the corridor and the printer area, furniture is mostly fixed and objects are less moveable. As a result, these areas were less susceptible to variations caused by human activity

(a) Minnie. (b) Camera. (c) Dumbo.

Fig. 1. Devices employed in the acquisition: the two mobile robot platforms "Minnie" (a) and "Dumbo" (c) as well as the standard camera on a tripod (b).

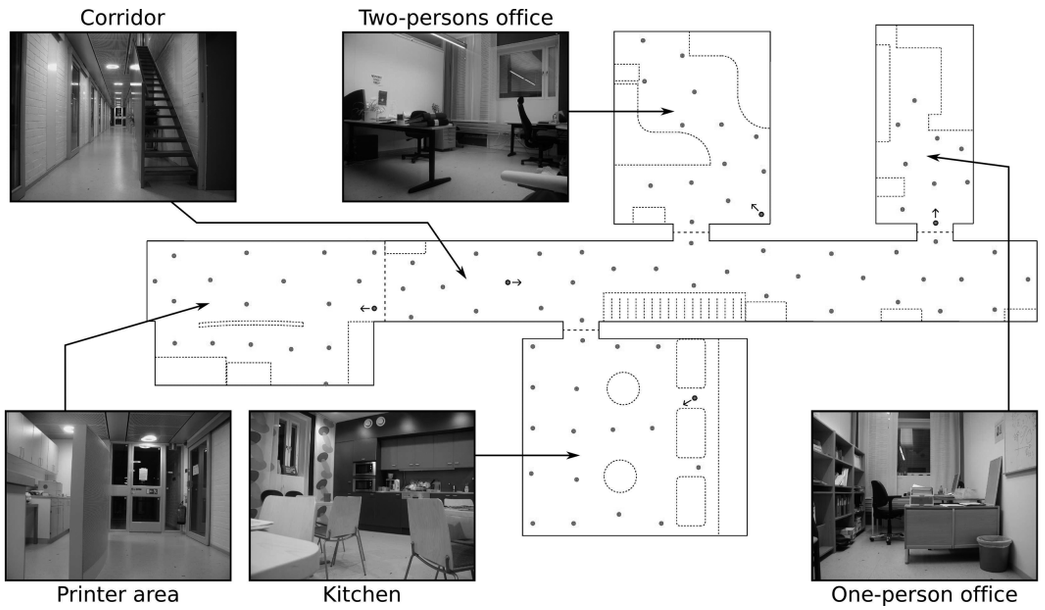


Fig. 2. A general map of the part of the office environment that was imaged during acquisition of the INDECS and IDOL databases. Boundaries between the five rooms were marked with dashed lines. Dotted lines were used to draw an approximate outline of furniture. Moreover, the location of points at which the tripod was placed while recording the INDECS database were marked. The pictures are taken from the database and show the interiors of the five rooms. The small arrows were used to indicate the viewpoints at which the presented pictures were taken.

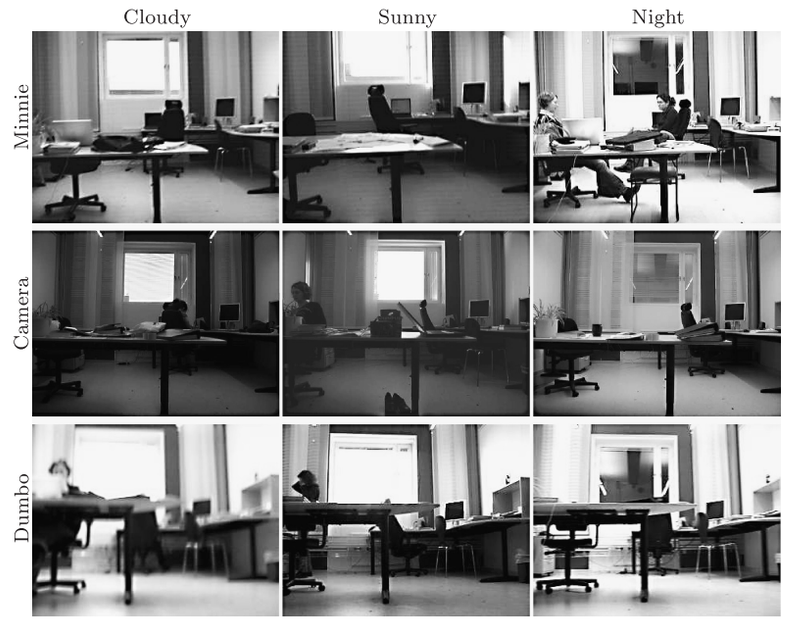
in comparison with the kitchen or the offices, where furniture (e.g. chairs) is relocated more often and objects (e.g. cups, laptops, etc.) are frequently moved.

The rooms are physically separated by sliding glass doors. The printer area is an exception and was treated as a separate place only due to its different functionality (the border between the corridor and the printer area was arbitrarily defined). The laboratory contains additional rooms which were not taken into consideration while creating the database. However, because of the glass door, other parts of the environment can still be visible in the images. Examples of pictures showing the interior of each room as well as a general map of the environment are presented in Fig. 2.

As already mentioned, the visual data were acquired with three different devices. In each case, the appearance of the rooms was captured under three different illumination and weather conditions: in cloudy weather (natural and artificial light), in sunny weather (direct natural light dominates) and at night (only artificial light). Since all the rooms have windows, the influence of natural illumination was significant. The image acquisition was spread over a period of time of 3 months, for the INDECS database, and over 2 weeks for the IDOL database. Additionally, the INDECS database was acquired 10 months before the experiments with the robots. In this way, we captured the visual variability that occurs in realistic environments due to varying illumination and natural

(a) Corridor.

(b) Kitchen.



(c) Two-person office.

Fig. 3. Example pictures taken from the INDECS and IDOL databases acquired with the camera and the two robot platforms under various illumination conditions. The pictures show the influence of illumination (especially (a) and (c)) and illustrate the differences between images acquired in a cluttered environment using different devices (b). Additional variability caused by natural activities in the rooms is also apparent (presence of people, relocated objects and furniture).

activities in the rooms. Fig. 3 presents a comparison of images taken under different illumination conditions and using various devices.

4.2. The INDECS database

The INDECS database consists of pictures of the environment described above, gathered from different viewpoints using a standard camera mounted on a tripod. We marked several points in each room (approximately one meter apart) where we positioned the camera for each acquisition. The rough positions of all points are shown on the map in Fig. 2. The number of points changed with the dimension of the room, from a minimum of 9 for the one-person office to a maximum of 32 for the corridor. At each location we acquired 12 images, one every 30°, even when the tripod was located very close to a wall or furniture. Examples of images taken at the same location and from several angles are presented in Fig. 4. Images were acquired using an Olympus C-3030ZOOM digital camera and the height of the tripod was constant and equal to 76 cm. All images in the INDECS database were acquired with a resolution of 1024×768 pixels, the auto-exposure mode enabled, flash disabled, the zoom set to wide-angle mode and the auto-focus enabled. In this paper, the INDECS images were subsampled to 512×386 before being used in the experiments. The images were labeled according to the position of the point where the acquisition happened. As a consequence, images taken, for example, from the corridor but looking into a room are labeled as the corridor. The images were acquired across a time span of 3 months and under varying illumination conditions (sunny, cloudy and night). For each illumination setting, we captured one full set of images. In total, there are 3264 images (324 for the one-person office, 492 for the two-person office, 648 each for the kitchen and the printer area and 1152 for the corridor) in the INDECS database.

4.3. The IDOL database

The IDOL database was acquired using cameras on two mobile robot platforms. Both robots, the PeopleBot Minnie and the PowerBot Dumbo, were equipped with a pan-tilt-zoom Canon VC-C4 camera, a SICK laser range finder and wheel encoders. However,

Fig. 4. Pictures from the INDECS database taken from several angles at the same location in the two-person office.

as it can be seen from Fig. 1, the cameras were mounted at different heights. On Minnie, the camera was 98 cm above the floor, whereas on Dumbo it was 36 cm. Furthermore, the camera on Dumbo was tilted up approximately 13°, to reduce the amount of floor captured in the images. The selected positions of the cameras result in different characteristics of the environment being captured in the images. Due to the low placement of the camera on Dumbo, the captured images are less susceptible to variations introduced by human activity in the environment and direct sunlight coming through the windows. At the same time, the camera on Minnie was able to capture the appearance of objects located on the desks and provide more information about the semantics of a place. All images were acquired with a resolution of 320 × 240 pixels, with the zoom fixed to wide angle (roughly 45° horizontal and 35° vertical field of view), and the auto-exposure and the auto-focus modes enabled.

We followed the same procedure during image acquisition with both robot platforms. Each robot was manually driven (average speed around 0.3–0.35 m/s) through each of the five rooms while continuously acquiring images at the rate of five frames per second. The path was roughly planned so that the robots could capture the visual appearance of all the rooms. For the different illumination conditions (sunny, cloudy, night), the acquisition procedure was performed twice, resulting in two image sequences acquired one after another giving a total of six sequences for each robot platform across a span of over 2 weeks. Each of the image sequences in the database is accompanied by laser scans and odometry data. Due to the manual control, the path of the robot was slightly different for every sequence. Examples of paths are presented in Figs. 7–9. Each image sequence consists of 1000–1300 frames. To automate the process of labeling the images for the supervision, the robot pose was estimated during the acquisition process using a laser-based localization method [47]. Again, each image was labeled as belonging to one of the five rooms based on the position from where it was taken.

4.4. Acquisition results

Examples illustrating the properties of images that can be found in both databases are given in Fig. 3. First of all, we can observe a significant influence of illumination. The appearance of the rooms is affected by highlights, shadows and reflections, especially in the case of strong direct sunlight. Moreover, the fact that the auto-exposure mode was on, resulted in a lower contrast in the informative parts of images, when the camera was directed towards a bright window in sunny weather. At the same time, the conditions observed during cloudy weather were much more stable and could be seen as intermediate between those during sunny weather and at night. A second important type of variability was introduced by the human presence and activities. In some cases, people partially occluded the view. Furthermore, the fact that the environment was observed for some time allowed to capture different configurations of furniture or objects placed on the desks or kitchen tables. The fact that objects could be observed in the images makes it possible to use the database in more complex scenarios where place recognition and object recognition complement each other, e.g. by contextual priming [13,40] (especially in the case of the high resolution images in the INDECS database). Finally, we can compare the images acquired using the three different devices. We see that each device captures different aspects of the same environment, mainly due to the variations in viewpoints caused by the different heights of the cameras. The influence of viewpoint is substantial, especially for cluttered scenes, when the camera was close to the furniture.

For both databases, the environment was observed from multiple viewpoints. For INDECS, the viewpoints are stable over different weather conditions, but the appearance of the rooms is almost fully captured as the images were taken in 12 directions. In the case of IDOL, we observe changes in the viewpoint due to manual control of the robot, but since the robot was driven in a particular direction, parts of the environment might not be observed. As previously mentioned, labeling was based on the position of the camera rather than contents of the images, and acquisition was performed even close to walls or furniture. As a result, both databases contain difficult cases, where the contents of the image is either non-informative or is weakly associated with the label.

To summarize, despite the fact that the acquisition was performed in a relatively small environment (consisting of five different rooms), there are several types of variability captured which pose a challenge to a recognition system. These range from different acquisition conditions to large viewpoint variations across the devices. Moreover, the acquisition procedure was carefully designed, and each single dataset offers different, but usually well-specified, type of variability. As a result, the influence of different factors on the accuracy of the system can be isolated and precisely measured. The relatively small environment does not allow for concluding that a system evaluated on the data will offer similar absolute performance in a different environment. However, since the data capture the influence of a large amount of variations on the appearance of a standard office environment, we can expect that an algorithm robust to those variations will be robust to similar types of variations within other indoor office environments.

5. Baseline evaluation

This section presents the visual place recognition system with which we assessed the INDECS and IDOL databases. We applied a fully supervised, appearance-based method. It assumes that each room is represented, during training, by a collection of images capturing its visual appearance under different viewpoints, at a given time and illumination. During testing, the algorithm is shown images of the same rooms, acquired under roughly similar viewpoints but possibly under different illumination conditions and after some time (where the time range goes from some minutes to several months). The goal is to recognize correctly each single image seen by the system. The method is based on a large-margin discriminative classifier, namely SVMs [27] and two different image representations. We use global and local image features, and we combine them with SVMs through specialized kernels. As a result, the recognition process always consists of two steps: feature extraction and classification.

In the rest of this section, we first motivate the decision to provide a baseline evaluation with the presented datasets (Section 5.1). Then, we describe the employed image representations (Section 5.2) and the classifier (Section 5.3). Finally, we explain the procedure followed in our experiments (Section 5.4).

5.1. Motivation

An important component when providing the community with a new collection of data is to give a quantitative measure of how hard the database is. Benchmark databases have become a very popular tool in several research communities during the last years [19,43] because they provide at the same time an instrument to develop new state-of-the-art algorithms, and a way to call attention on a research topic. When a database is used for developing a new algorithm, it is extremely useful to be able to compare the obtained results with those obtained by some other established technique: this permits to understand what are the advantages of the new method over existing approaches. At the same time, presenting a new corpus together with a baseline evaluation helps the community to quickly identify the open challenges of the problem and therefore concentrate on their research efforts. While often the baseline evaluation consists of a newly developed method, very often it is a well known, off the shelf solution: again, the goal of a baseline evaluation is not to present a new theory, but to provide a quantitative evaluation of how challenging the new dataset is, coupled with a well-defined experimental protocol.

The computer vision community has been traditionally very open to the introduction of publicly available databases [19,43] and associated benchmark challenges [20]. These two tools, combined together, have heavily contributed to set the research agenda of the last years. The robotics community has recently started to acknowledge the value and power of such collections, as it is witnessed by several successful benchmark evaluations [48,49].

5.2. Feature extraction

The feature extraction step aims at providing a representation of the input data that minimize the within-class variability while at the same time maximizing the between-class variability. Additionally, this representation is usually more compact than raw input data and therefore allows us to reduce the computational load imposed by the classification process. Features can be derived from the whole image (global features) or can be computed locally, based on its salient parts (local features).

As environments can be described differently, depending on the considered scale, scale–space theory appears as a suitable framework for deriving effective representations here. Following this intuition, we chose to use two scale–space theory-based features, one global (composed receptive field histograms, CRFH [26]) and one local (scale invariant feature transform, SIFT [25]). The rest of the section describes briefly the two approaches.

5.2.1. Global features: Compose receptive field histograms

CRFH is a multi-dimensional statistical representation of the occurrence of responses of several image descriptors applied to the image. This idea is illustrated in Fig. 5. Each dimension corresponds to one descriptor and the cells of the histogram count the pixels sharing similar responses of all descriptors. This approach allows us to capture various properties of the image as well as relations that occur between them.

Multi-dimensional histograms can be extremely memory consuming and computationally expensive if the number of dimensions grows. For example, a 9-dimensional histogram with 16 quantization levels per dimension contains approximately 7×10^{10} cells. In [26], Linde and Lindeberg suggest to exploit the fact that most of the cells are usually empty and to store only those that are non-zero. The histogram can be stored in a sparse form as an array $[(c_1, v_1), (c_2, v_2), \ldots, (c_n, v_n)]$, where c_i denotes the index of the cell containing the non-zero value v_i. This representation allows us not only to reduce the amount of memory required, but also to perform operations such as histogram accumulation and comparison efficiently. For our experiments, we built multi-dimensional histograms using combinations of several image descriptors, applied to the scale–space representation at various scales, namely first- and second-order Gaussian derivatives, gradient magnitude, Laplacian and Hessian determinant applied to both intensity and color channels.

5.2.2. Local features: Scale invariant feature transform

The idea behind *local features* is to represent the appearance of an image only around a set of characteristic points known as the *interest points*. The similarity between two images is then measured by solving the correspondence problem. Local features are known to be robust to occlusions, as the absence of some interest points does not affect the features extracted from other local patches.

The process of local feature extraction consists of two stages: *interest point detection* and *description*. The interest point detector identifies a set of characteristic points in the image that could be re-detected even in spite of various transformations (e.g. rotation and scaling) and variations under illumination conditions. The role of the descriptor is to extract robust features from the local patches located at the detected points.

In this paper, we used the scale, rotation and translation invariant Harris–Laplace detector [50] and the commonly used SIFT descriptor [25]. Comparisons of local descriptors and interest point detectors, presented in [51], show that both algorithms are highly reliable. Moreover, the SIFT descriptor has shown to perform well for object classification ([52]) and mobile robot localization ([33,29]).

5.3. Classification: Support vector machines

The choice of the classifier is the second key ingredient for an effective visual place recognition system. In this paper, we chose SVMs based on their state-of-the-art performances in several visual recognition domains [53–55]. The rest of this section reviews briefly the theory behind the algorithm and describes our choices for the kernel function. We refer the readers to [27] for a thorough introduction to the subject.

5.3.1. Linear SVM

Consider the problem of separating a set of training data $(\boldsymbol{x}_1, y_1), \ldots (\boldsymbol{x}_m, y_m)$ into two classes, where $\boldsymbol{x}_i \in \Re^N$ is a feature vector and $y_i \in \{-1, +1\}$ its class label. Assuming that the two classes can be separated by a hyperplane $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$, the optimal hyperplane will be the one with maximum distance to the closest points in the training set. The optimal values for $\boldsymbol{w}$ and b can be found by solving a constrained minimization problem via Lagrange multipliers, resulting in a classification function

$$f(\boldsymbol{x}) = \operatorname{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x_i} \cdot \boldsymbol{x} + b\right), \tag{1}$$

where α_i and b can be found efficiently using the sequential minimal optimization (SMO, [56]) algorithm. The $\boldsymbol{x}_i$ with $\alpha_i \neq 0$ are called *support vectors*.

5.3.2. Nonlinear SVM and kernel functions

To obtain a nonlinear classifier, one maps the data from the input space $\Re^N$ to a higher-dimensional feature space $\mathcal{H}$ by $\boldsymbol{x} \rightarrow \Phi(\boldsymbol{x}) \in \mathcal{H}$, such that the mapped data points of the two classes are linearly separable in the feature space. Assuming that there exists a kernel function K such that $K(\boldsymbol{x}, \boldsymbol{y}) = \Phi(\boldsymbol{x}) \cdot \Phi(\boldsymbol{y})$, a nonlinear SVM can be constructed by replacing the inner product $\boldsymbol{x_i} \cdot \boldsymbol{x}$ by the

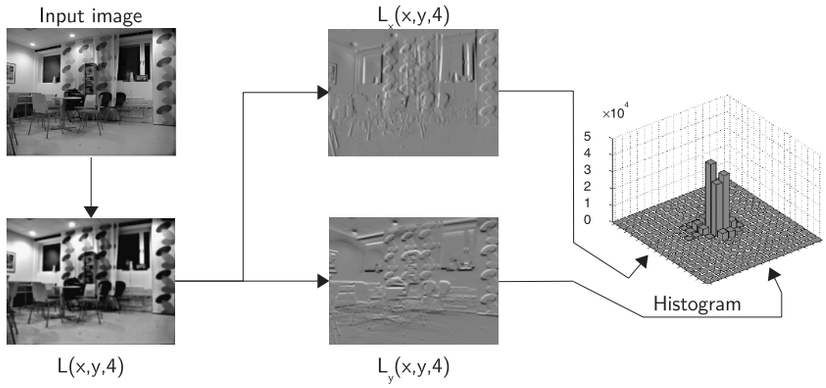


Fig. 5. The process of generating multi-dimensional receptive field histograms shown in the example of the first-order derivatives computed at the same scale $t = 4$ from the illumination channel.

kernel function $K(\boldsymbol{x_i}, \boldsymbol{x})$ in Eq. (1). This corresponds to constructing an optimal separating hyperplane in the feature space.

The choice of the kernel function is a key ingredient for the good performance of SVMs; based on results reported in the literature, we chose in this paper the χ^2 *kernel* [57] for global features and the *match kernel* [58] for local features.

The χ^2 kernel belongs to the family of exponential kernels, and is given by

$$K(\boldsymbol{x}, \boldsymbol{y}) = \exp\left\{-\gamma\, \chi^2(\boldsymbol{x}, \boldsymbol{y})\right\}, \qquad \chi^2(\boldsymbol{x}, \boldsymbol{y}) = \sum_i \frac{\|x_i - y_i\|^2}{\|x_i + y_i\|}. \quad (2)$$

The match kernel is given by [58]

$$K(\boldsymbol{L}_h, \boldsymbol{L}_k) = \frac{1}{n_h} \sum_{j_h=1}^{n_h} \max_{j_k=1,\ldots,n_k} \left\{K_l(\boldsymbol{L}_h^{j_h}, \boldsymbol{L}_k^{j_k})\right\}, \quad (3)$$

where $\boldsymbol{L}_h$ and $\boldsymbol{L}_k$ are local feature sets and $\boldsymbol{L}_h^{j_h}$ and $\boldsymbol{L}_k^{j_k}$ are two single local features. The sum is always calculated over the smaller set of local features and only some fixed amount of best matches is considered in order to exclude outliers. The local feature similarity kernel K_l can be any Mercer kernel. We used the RBF kernel based on the Euclidean distance for the SIFT features:

$$K_l(\boldsymbol{L}_h^{j_h}, \boldsymbol{L}_k^{j_k}) = \exp\left\{-\gamma \|\boldsymbol{L}_h^{j_h} - \boldsymbol{L}_k^{j_k}\|^2\right\}. \quad (4)$$

The match kernel was introduced in [58], and despite the claim in the paper, it is not a Mercer kernel [59]. Still, it can be shown that it statistically approximates a Mercer kernel in a way that makes it a suitable kernel for visual applications [59]. On the basis of this finding, and of its reported effectiveness for object categorization [53], we will use it here.

5.3.3. Multi-class SVM

The extension of SVM to multi-class problems can be done mainly in two ways:

- *One-vs-all strategy*. If M is the number of classes, M SVMs are trained, each separating a single class from all remaining classes. The decision is then based on the distance of the classified sample to each hyperplane and the final output is the class corresponding to the hyperplane for which the distance is largest.
- *One-vs-one strategy*. In this case, $M(M-1)/2$ two-class machines are trained for each pair of classes. The final decision can then be taken in different ways, based on the $M(M-1)/2$ outputs. A popular choice is to consider as output of each classifier the class label and count votes for each class; the test image is then assigned to the class that received more votes. Another alternative is to use signed distance from the hyperplane and sum distances for each class. Other solutions based on the idea to arrange the pairwise classifiers in trees, where each tree node represents an SVM, have also been proposed [60,27].

In this paper, for efficiency reasons, we will use the pairwise approach and the voting-based method, which we found to constantly outperform the second variant in preliminary experiments (the complexity of the SVM training algorithm is approximately $O(n^2)$ and smaller training subsets of the binary classifiers make the training procedure faster).

5.4. Experimental setup

We conducted four series of experiments in order to assess thoroughly the INDECS and IDOL databases. For each series of experiments, we evaluated the performance of both local and global image representations. We divided the databases into several subsets with respect to the illumination conditions that prevailed during acquisition and the device employed. For the INDECS database, we considered three image sets, one for each illumination setting (cloudy, night, sunny). Since the IDOL database consists of 12 image sequences, we used each full sequence as a separate set. The system was always trained in a supervised fashion on one, two or three datasets and tested on a fourth different set. In order to test the limits of the underlying visual recognition algorithm, we considered each image in the test set separately, and as a final measure of performance, we used the percentage of properly recognized images. As the number of acquired images varied across rooms, the performance obtained for each place was considered separately during the experiments. The final classification rate was then computed as the average between all the rooms' results. This procedure ensures that performance on each place contributes equally to the overall result, thus avoiding the biases towards areas with many acquired images, such as the corridor.

We started with a set of reference experiments, assessing the data acquired under stable illumination. To achieve this, for training and testing we used datasets acquired with the same device and under similar conditions. Next, we increased the difficulty of the problem and tested the robustness of the system to changing illumination conditions as well as to other variations that may occur in real-world environments. Training and recognition were in this case performed on datasets consisting of images captured under different illumination settings and usually on different days. The third set of experiments aimed to reveal whether a model trained on an image set acquired with one device can be useful for solving localization problem with a different

device (and usually after some time). Finally, we checked whether the robustness of the recognition algorithm can be increased by providing additional training data capturing a wider spectrum of visual variability. For that, we trained the system on two or three image sets gathered under different illumination conditions. Additionally, before carrying out the benchmarks described above, we conducted a set of preliminary experiments in order to select proper kernel functions and feature extractor parameters. All the results obtained on these experiments are reported in Section 6.

For all experiments, we used our extended implementation of SVMs based on the *libsvm* software [61]. We set the value of the error penalty C to be equal to 100 and we determined the kernel parameters via cross-validation.

6. Experimental results

This section reports the results of the baseline evaluation of the INDECS and IDOL databases, according to the procedure described in Section 5.4. We present the results in consecutive subsections, and we give a brief summary in Section 6.5.

As described in Section 5.4, before performing the actual benchmark, we ran a set of preliminary experiments on the INDECS database, mainly using the global features (CRFH). We evaluated the performance of the multi-dimensional histograms built from a wide variety of combinations of global image descriptors listed in Section 5.2 for several scale levels and numbers of histogram bins per dimension. A comprehensive report on the obtained results can be found in [62]. The experiments revealed that the most valuable global features can be extracted using non-isotropic, derivative-based descriptors, and that chromatic cues are more susceptible to illumination variations. As a result, here we used composed receptive field histograms of six dimensions with 28 bins per dimension, computed from second-order normalized Gaussian derivative filters, applied to the illumination channel at two scales. The scale levels were different for the experiments with IDOL ($\sigma =$ 1 and 4) and with INDECS ($\sigma = 2$ and 8). This was motivated by the fact that the cameras mounted on the robots obtained images of lower quality, and their movement introduced additional distortions.

6.1. Stable illumination conditions

In order to evaluate our method under stable illumination conditions, we trained and tested the system on pairs of image sequences taken from the IDOL database acquired one after the other using the same robot. As mentioned previously, we analyzed the performance of both global (CRFH) and local (SIFT) image descriptors. We did not use the INDECS database for these experiments since only one set of data for each illumination setting was available. Although the illumination conditions for both training and test images were in this case very similar, the algorithm had to tackle other kinds of variability such as viewpoint changes (caused mainly by the manual control of the robot) and the presence/absence of people. The results of the performed experiments are presented in Fig. 6a, c for CRFH and in Fig. 6b, d for SIFT. For each platform and type of illumination conditions used for training, the first bar presents an average classification rate over the two possible permutations of the image sequences in the training and test sets.[1] On average, the system classified properly 95.5% of the images acquired with Minnie and 97.3% of images acquired with Dumbo when global features were used. When local features were applied, the average recognition rates were slightly lower and equal to 94.4% and 94.9%, respectively.

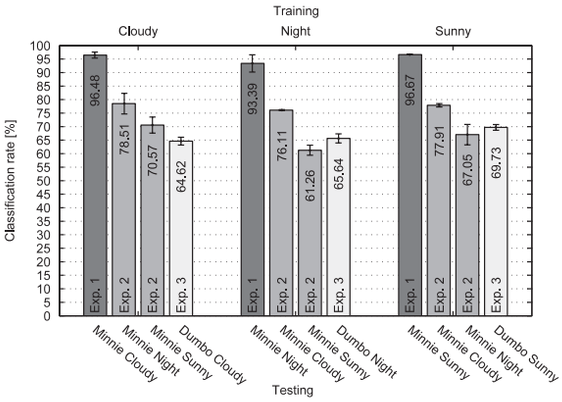
[1] Training on the first sequence, testing on the second sequence, and vice versa.

Detailed results for two experiments conducted on data captured with each of the platforms are shown in Fig. 7. The figure presents maps of the environment with plotted paths of the robot during acquisition of the training and test sequences used during each of the experiments. Moreover, the symbols used to draw the test path indicate the results of recognition performed using image acquired at each location. Each experiment started at the point marked with the label "Start" and the arrows show the direction of driving. The position of the furniture (plotted with gray line) is approximate and sometimes slightly varied between the experiments. It can be observed that the errors are usually not a result of viewpoint variations (compare the training and test paths in the kitchen, especially in Fig. 7c, d) and mostly occur near the borders of the rooms. This can be explained by the relatively narrow field of view of the cameras as well as the fact that the images were not labeled according to their content but to the position of the robot at the time of acquisition. Since these experiments were conducted with the sequences captured under similar conditions, we treat them as a reference for other results.

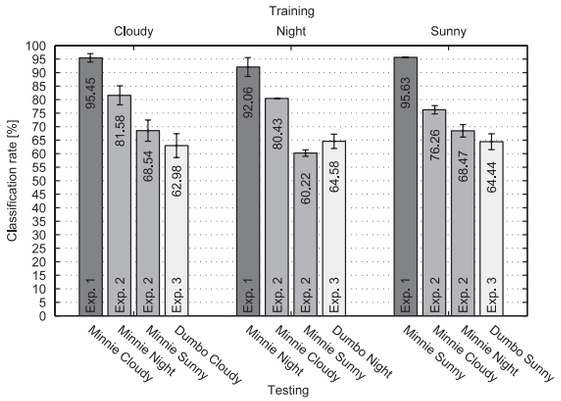
6.2. Varying illumination conditions

We also conducted a series of experiments aiming to test the robustness of our method to changing illumination conditions as well as to other variations caused by normal activities in the rooms. The experiments were conducted on both INDECS and IDOL databases. As with the previous experiments, the same device was used for both training and testing. This time, however, the selected training and testing datasets consisted of images acquired under different illumination conditions and usually on different days. Fig. 6a–d show average results of the experiments with the image sequences from the IDOL database acquired with both robots for each permutation of the illumination conditions used for training and testing and both image representations (the two middle bars for each figure and type of training conditions). The presented classification rates obtained on the IDOL database were always averaged over two experiments with different image sequences. Fig. 6e,f gives corresponding results obtained on the INDECS database.

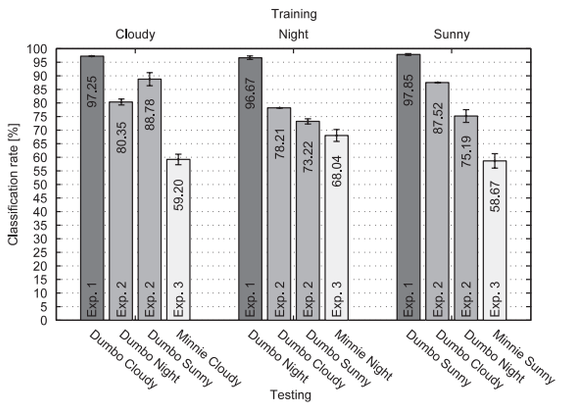
We see that, in general, the system performs best when trained on the images acquired in cloudy weather. The explanation for this is straightforward: the illumination conditions on a cloudy day can be seen as intermediate between those at night (only artificial light) and on a sunny day (direct natural light dominates). In such case, the average classification rate computed over two testing illumination conditions (sunny and night) for both CRFH and SIFT was equal to 84.6% and 87.3% for Dumbo, 74.5% and 75.1% for Minnie, and 81.3% and 76.4% for the INDECS database. In general, local features performed slightly better than the global features (in average 71.9% vs. 72.6% for Minnie and 80.5% vs. 83.2% for Dumbo), although it was usually not the case for the INDECS database (in average 75.9% vs. 72.5%). Fig. 8 presents detailed results for two example runs and both feature types. The errors occurred mainly for the same reasons as in the previous experiments and additionally in places heavily affected by the natural light, e.g. when the camera was directed towards a bright window or, in particular, large glass door in the printer area. In such cases, the automatic exposure system with which all the cameras were equipped caused the pictures to darken. Minnie was more susceptible to this phenomenon due to the higher position of its camera.



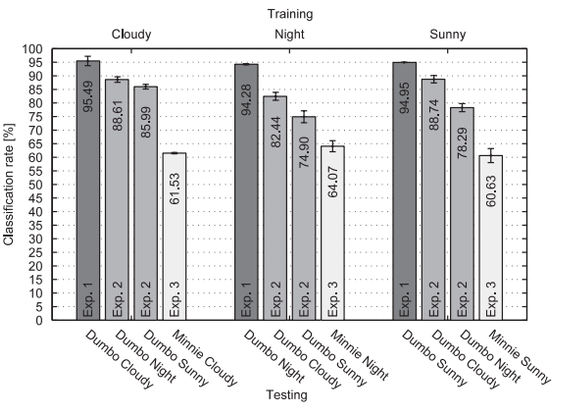
(a) Training on global features (*CRFH*) extracted from images acquired with it Minnie.



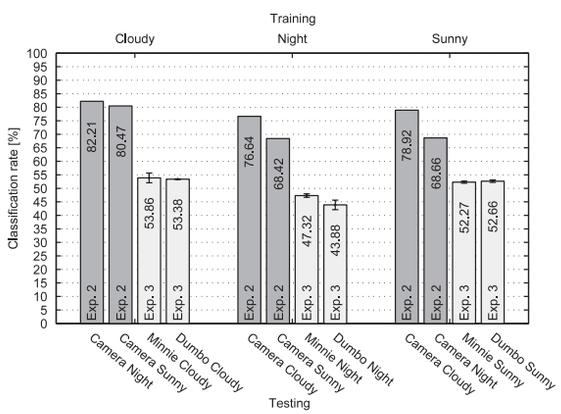
(b) Training on local features (*SIFT*) extracted from images acquired with it Minnie.



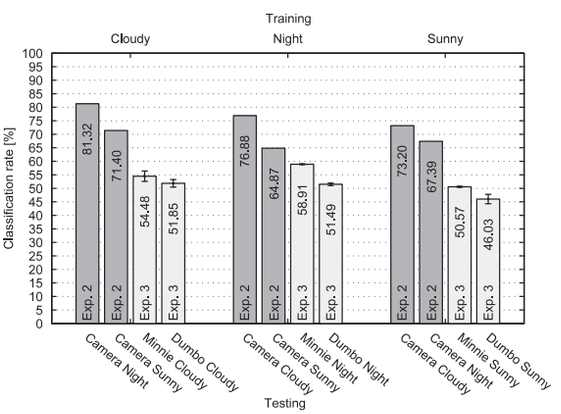
(c) Training on global features (*CRFH*) extracted from images acquired with it Dumbo.



(d) Training on local features (*SIFT*) extracted from images acquired with it Dumbo.



(e) Training on global features (*CRFH*) extracted from images acquired with the it Standard camera.

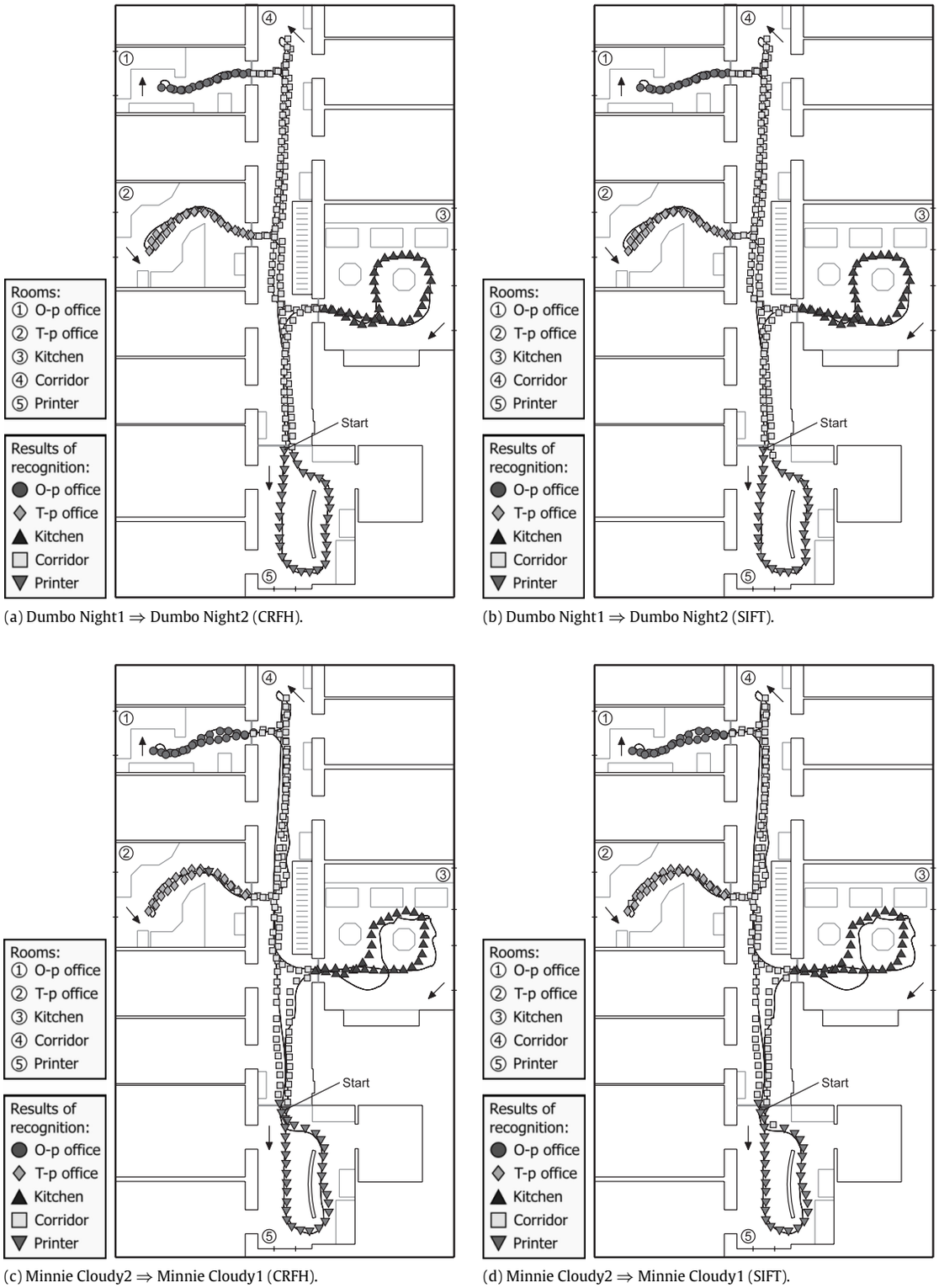


(f) Training on local features (SIFT) extracted from images acquired with it Standard camera.

Fig. 6. Average results of the first three experiments on the IDOL and INDECS databases with both image representations. In each figure, the results are grouped according to the type of illumination conditions under which the training images were acquired. The bottom axes indicate the platform and illumination conditions used for testing. The uncertainties are given as one standard deviation.

6.3. Recognition across platforms

The third set of experiments was designed to test the portability of the acquired model across different platforms. For that purpose, we trained and tested the system on image sets acquired under similar illumination conditions using different devices. We started with the experiments on image sequences from the IDOL database. We trained the system on the images acquired using either Minnie or Dumbo and tested with the images captured with the other robot. We conducted the experiments for all illumination conditions and both image representations. The main difference between the platforms from the point of view of our experiments lies in the height at which the cameras are mounted (98 cm for Minnie and 36 cm for Dumbo). The results presented in Fig. 6a–d indicate that our method was still able to classify correctly up to about 70% of images for CRFH and 65% of images for SIFT. There was no clear advantage of using one particular feature type. The system performed better when trained on the images captured with Minnie. This can be explained by the fact that the lower mounted camera on Dumbo provided less diagnostic information. It can also be observed from Fig. 9 that, in general, the additional errors occurred when the robot was positioned close to the walls



(a) Dumbo Night1 ⇒ Dumbo Night2 (CRFH).

(b) Dumbo Night1 ⇒ Dumbo Night2 (SIFT).

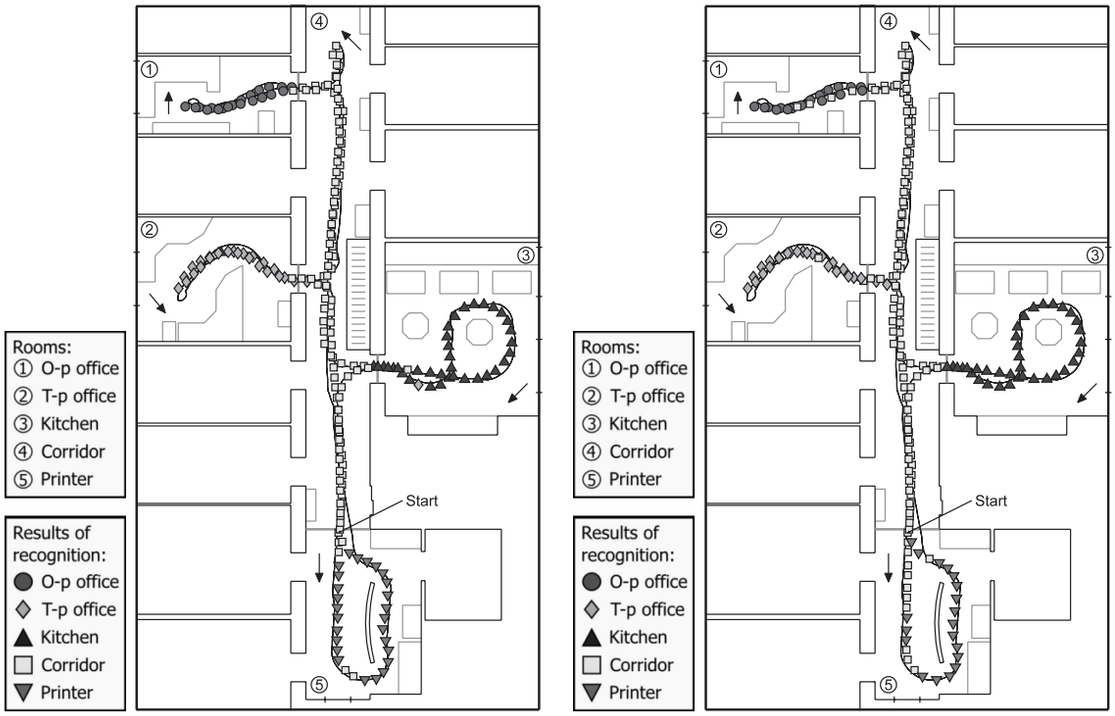
(c) Minnie Cloudy2 ⇒ Minnie Cloudy1 (CRFH).

(d) Minnie Cloudy2 ⇒ Minnie Cloudy1 (SIFT).

Fig. 7. Maps of the environment with plotted paths of the robot during acquisition of the training (black line) and test (points) sequences taken from the IDOL database and used during the experiments with *stable illumination conditions*. The shape of each point on the test path indicates the result of recognition.

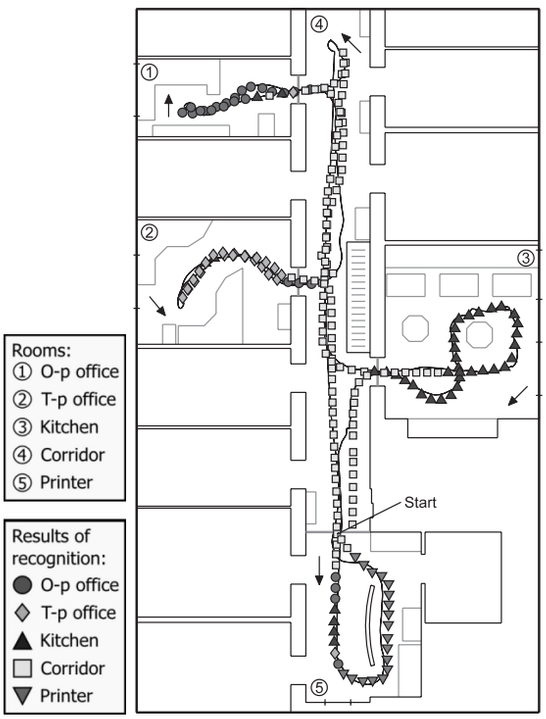
or furniture. In such cases, the height of the camera influenced the content of the images the most.

We followed a similar procedure using the INDECS database as a source of training data and different image sequences taken from the IDOL database for testing. It is important to note that the acquisition procedure differed in the case of both databases, and the INDECS database was gathered 10 months before the acquisition of IDOL. The points at which the pictures were taken were positioned approximately 1 m from each other and, in the case of the kitchen, covered different area of the room due to reorganization of the furniture. Consequently, the problem required that the algorithm was not only invariant to various acquisition techniques but also offered great robustness to large changes in the viewpoint and the appearance of the rooms

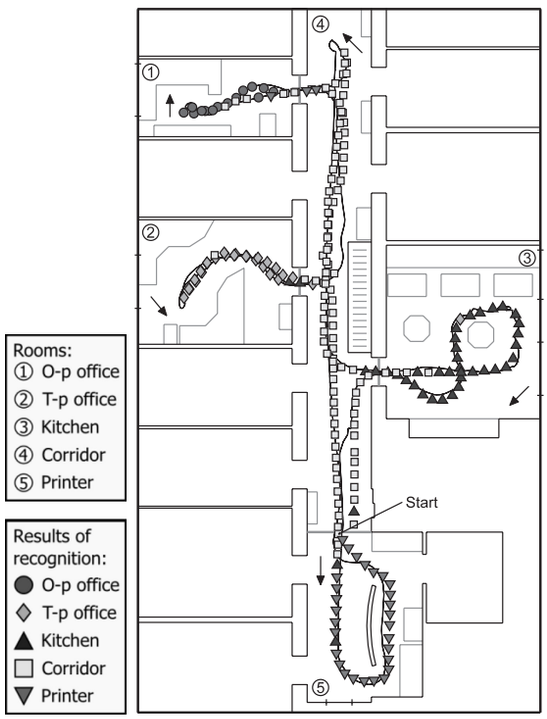


(a) Dumbo Cloudy2 ⇒ Dumbo Sunny2 (CRFH).

(b) Dumbo Cloudy2 ⇒ Dumbo Sunny2 (SIFT).



(c) Minnie Cloudy2 ⇒ Minnie Night2 (CRFH).



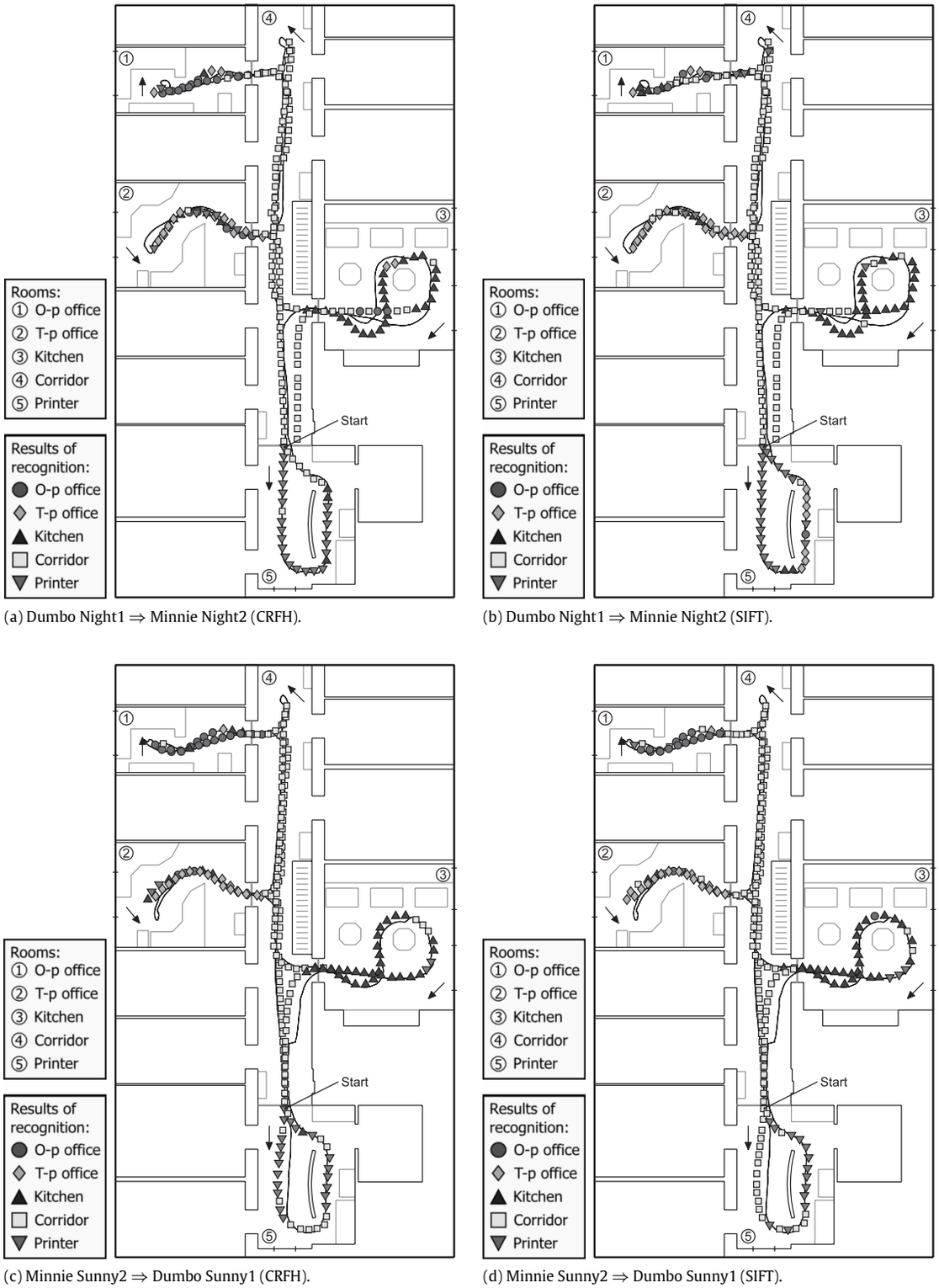
(d) Minnie Cloudy2 ⇒ Minnie Night2 (SIFT).

Fig. 8. Maps of the environment with plotted paths of the robot during acquisition of the training (black line) and test (points) sequences taken from the IDOL database and used during the experiments with *varying illumination conditions*. The shape of each point on the test path indicates the result of recognition.

introduced by long-time human activity. The experimental results are presented in Fig. 6e, f. We see that the algorithm obtains a recognition performance of about 50%. While this result is surely disappointing if compared to the 70% reported above, obtained for the two robot platforms, it is still quite remarkable considering the very high degree of variability between training and test data, and that results are significantly above chance (which in this case would be 20% as the datasets contain images acquired in five rooms).

6.4. Training-based robustness

The final series of experiments aimed at revealing whether the robustness of the recognition algorithm can be boosted by pro-



(a) Dumbo Night1 ⇒ Minnie Night2 (CRFH).

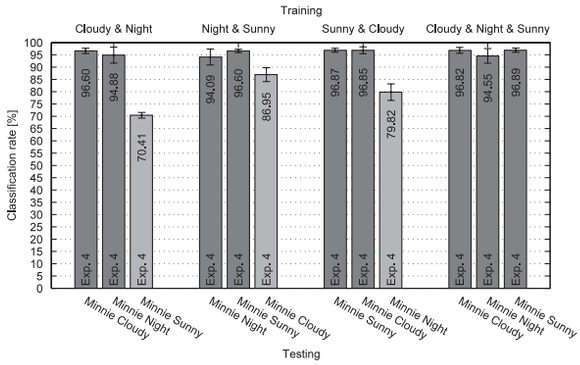
(b) Dumbo Night1 ⇒ Minnie Night2 (SIFT).

(c) Minnie Sunny2 ⇒ Dumbo Sunny1 (CRFH).

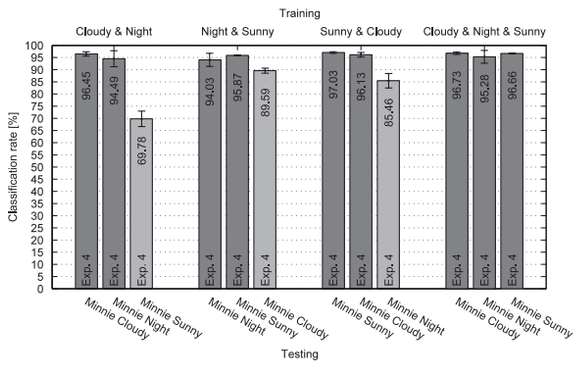
(d) Minnie Sunny2 ⇒ Dumbo Sunny1 (SIFT).

Fig. 9. Maps of the environment with plotted paths of the robot during acquisition of the training (black line) and test (points) sequences taken from the IDOL database and used during the experiments with *recognition across platforms*. The shape of each point on the test path indicates the result of recognition.

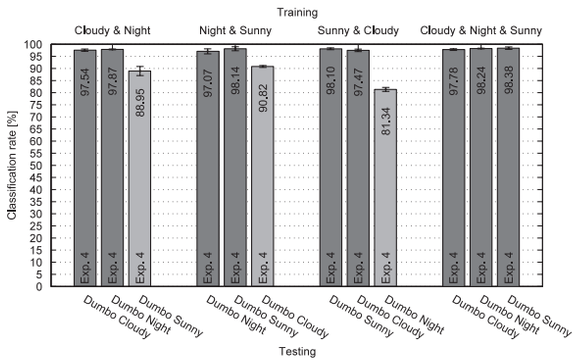
viding additional training data capturing a wider spectrum of visual variability that might occur in a real-world environment. In particular, we concentrated on invariance to changing illumination conditions as this is the kind of variability that a continuously running visual recognition system has to deal with every day. To achieve that, we trained the system on two or three image sequences from the IDOL database gathered under different illumination conditions, and we evaluated the recognition performance on another, fourth, image set. The obtained results for both platforms, all combinations of image sequences used for training as well as both CRFH and SIFT are presented in Fig. 10a–d. The darker bars indicate the results of experiments corresponding to those discussed in Section 6.1, when training was done on an image sequence acquired under conditions similar to those used for testing. The re-



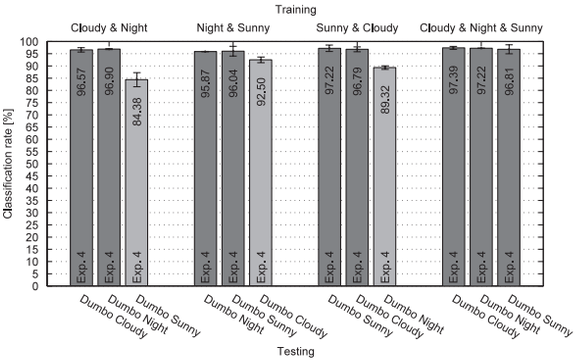
(a) Training on global features (*CRFH*) extracted from images acquired with *Minnie*.



(b) Training on global features (*SIFT*) extracted from images acquired with *Minnie*.



(c) Training on global features (*CRFH*) extracted from images acquired with *Dumbo*.



(d) Training on global features (*SIFT*) extracted from images acquired with *Dumbo*.

Fig. 10. Performance of the system trained on two or three image sequences acquired under different illumination conditions for both mobile platforms and image representations. The classification rates were averaged over all possible combinations of training and test sequences. The uncertainties are given as one standard deviation.

sults shown using the brighter bars can be compared with those of the experiments under varying illumination conditions analyzed in Section 6.2.

It is apparent that including images acquired under different conditions into the training set improves recognition accuracy. Although the algorithm has to incorporate much more information about each of the places into the model, the recognition accuracy for test sets acquired under similar conditions as those used for training is even greater than this obtained when each training sequence was used separately (as for the experiments discussed in Section 6.1). For example, the average recognition rate over all test sets and illumination settings for models trained on three sequences acquired using Dumbo was equal to 98.1% for CRFH and 97.1% for SIFT. At the same time, for the experiments with stable illumination conditions reported in Section 6.1 (see Fig. 6), we got only 97.3% and 94.9%. The same trend can be observed for sequences captured using Minnie. Concluding, the ability of the algorithm to handle large within-class variability is clearly not a limiting factor. It is important to note that the recognition rate for conditions which were not used during training is also greatly improved when more training data are provided. For example, if the system was trained using the images captured during sunny weather and at night using Minnie, the average classification rate for testing image sequence acquired with cloudy weather was equal to 86.95% for CRFH and 89.59% for SIFT. Consequently, the classification rate improved by 9.9% in the case of CRFH and 11.2% in case of SIFT for testing conditions not known during training, at the same time slightly improving the rates for testing conditions used also for training.

It has to be pointed out that due to the larger number of training images capturing different types of variability, the number of support vectors stored in the final model grows as well. In such case, the user pays the price of the recognition time and the memory requirements, which in the case of SVMs grow linearly with the number of support vectors.

6.5. Discussion

The results of the extensive experimental evaluation presented in this section indicate that our method is able to perform place recognition using standard visual sensors with high precision. It offers good robustness to changes in the illumination conditions as well as to additional variations introduced by the natural variability that occurs in real-world environments. At the same time, there is a difference in the performance of the system between the experiments under stable and varying conditions, indicating that there is room for improvement in this matter.

As the system is to be used on a robot platform, it must not only be accurate but also efficient. For this reason, we tried to provide the highest possible robustness using relatively small amount of training data acquired during only one run. We managed to achieve a recognition time of less than 200 ms per frame on a Pentium IV 2.6 GHz using the global image representation. The results reported in Section 6.4 indicate that it is possible to significantly improve the robustness by incorporating images acquired during two or three runs under different illumination conditions into one training set. However, the higher performance does not come without a price. Since the number of support vectors in such case even doubles, the recognition time increased by about 50 ms.

In all the experiments, we evaluated both global (CRFH) and local (SIFT) image descriptors. In general, we did not find any clear advantage of using one feature type over the other, and each representation has its strengths and weaknesses. The global features, however, clearly outperform SIFT in terms of efficiency since the matching process required in order to compare two sets

of local patches is computationally expensive. The efficiency of the solution based on local features could be improved by applying a more efficient matching algorithm (e.g. by using a pyramid match SVM kernel [63]) or faster interest point detector and more compact descriptor (e.g. SURF [37,34]). Since global and local representations capture different aspects of a scene, the robustness of the final solution can be further improved by integrating both cues as proposed in [14,18].

7. Summary

This paper discussed the need for standard benchmarking solutions for vision-based topological localization, with particular emphasis on visual place recognition. We defined and analyzed carefully the problem, and we specified the open challenges that need to be addressed by a realistic benchmark. We presented two new databases, acquired on the basis of this analysis. The first, the INDECS database, contains pictures captured with a standard camera mounted on a tripod. The second, the IDOL database, contains image sequences acquired using cameras mounted on two mobile robot platforms. The two databases were recorded within the same indoor office environment. They capture a wide spectrum of natural variations introduced by both changing illumination and human activity. Each database can be seen as a different approach to the problem; thus, they can be used to analyze different properties of a place recognition system.

We assessed both databases with a large set of baseline experiments, using a fully supervised visual place recognition system. The method employs a large-margin discriminative classifier and two different image representations: a local representation, based on SIFT features, and a global representation, consisting of multi-dimensional histograms of receptive fields. We conducted the experiments according to an experimental procedure designed to contain problems of varying complexity and exploit most of the variability captured in the datasets. The experimental procedure can be seen as a part of the benchmark proposed in this paper. We started from experiments performed under stable illumination settings. We then performed experiments testing the robustness of the algorithms to changing illumination and human activity. Finally, we conducted experiments with large viewpoint variations and different acquisition methods.

The reported results show that the method is able to recognize places with high precision when training and testing is performed within a relatively stable environment, or when enough training data are provided. At the same time, there is space for improvement in the robustness to illumination and large viewpoint variations. The database still poses a challenge to the system which should provide stable performance in the presence of variability usually observed in real-world environments.

Finally, the dependence between the overall performance of the system and the particular set of data becomes visible as the complexity of the problem grows. Moreover, different methods (in this case different image descriptors) perform differently for different types of variations. This emphasizes the need for an extensive experimental evaluation, on a common benchmark dataset, for the comparison of different approaches. When realistic datasets are available, more extensive evaluation can be conducted as the data can be reused, fully exploited, and less effort is required for acquisition and annotation. The authors believe that benchmarking solutions, such as the one presented in this paper, will make an impact on the research on visual place recognition and topological localization as was the case for other localization and visual recognition problems.

Acknowledgments

This work was sponsored by the SSF through its Centre for Autonomous Systems (CAS), the EU integrated projects CoSy FP6-004250-IP, CogX ICT-215181 and DIRAC IST-027787 and the Swedish Research Council contract 2005-3600-Complex. The support is gratefully acknowledged.

References

[1] M. Jogan, A. Leonardis, Robust localization using an omnidirectional appearance-based subspace model of environment, Robotics and Autonomous Systems 45 (1) (2003) 51–72.
[2] M. Dissanayake, P. Newman, S. Clark, H. Durrant-Whyte, M. Csorba, A solution to the simultaneous localization and map building (SLAM) problem, IEEE Transactions on Robotics and Automation 17 (3) (2001) 229–241.
[3] J. Wolf, W. Burgard, H. Burkhardt, Robust vision-based localization by combining an image retrieval system with monte carlo localization, IEEE Transactions on Robotics 21 (2) (2005) 208–216.
[4] I. Ulrich, I. Nourbakhsh, Appearance-based place recognition for topological localization, in: Proceedings of the IEEE International Conference on Robotics and Automation, ICRA'00, San Francisco, CA, USA, 2000.
[5] M.M. Ullah, A. Pronobis, B. Caputo, J. Luo, P. Jensfelt, H.I. Christensen, Towards robust place recognition for robot localization, in: Proceedings of the 2008 IEEE International Conference on Robotics and Automation, ICRA'08, Pasadena, CA, USA, May 2008.
[6] M. Cummins, P. Newman, FAB-MAP: Probabilistic localization and mapping in the space of appearance, The International Journal of Robotics Research 27 (6) (2008) 647–665.
[7] S. Thrun, Learning metric-topological maps for indoor mobile robot navigation, Artificial Intelligence 1999 (1) (1998).
[8] E. Brunskill, T. Kollar, N. Roy, Topological mapping using spectral clustering and classification, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS'07, San Diego, October 2007.
[9] H. Zender, O.M. Mozos, P. Jensfelt, G.-J.M. Kruijff, W. Burgard, Conceptual spatial representations for indoor mobile robots, Robotics and Autonomous Systems 56 (6) (2008) 493–502.
[10] I. Nourbakhsh, R. Powers, S. Birchfield, Dervish: An office navigation robot, AI Magazine 16 (2) (1995) 53–60.
[11] O. Mozos, R. Triebel, P. Jensfelt, A. Rottmann, W. Burgard, Supervised semantic labeling of places using information extracted from sensor data, Robotics and Autonomous Systems 55 (5) (2007).
[12] B. Kuipers, P. Beeson, Bootstrap learning for place recognition, in: Proceedings of the 18th National Conference on Artificial Intelligence, AAAI'02, 2002.
[13] A. Torralba, K.P. Murphy, W.T. Freeman, M.A. Rubin, Context-based vision system for place and object recognition, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV'03, 2003, Nice, France.
[14] A. Pronobis, B. Caputo, Confidence-based cue integration for visual place recognition, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS'07, San Diego, CA, USA, October 2007.
[15] C. Siagian, L. Itti, Biologically-inspired robotics vision monte-carlo localization in the outdoor environment, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS'07, San Diego, CA, USA, October 2007.
[16] D. Kortenkamp, T. Weymouth, Topological mapping for mobile robots using a combination of sonar and vision sensing, in: Proceedings of the 12th National Conference on Artificial Intelligence, AAAI'94, Seattle, Washington, USA, 1994.
[17] A. Tapus, R. Siegwart, Incremental robot mapping with fingerprints of places, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS'05, Edmonton, Alberta, Canada, August 2005.
[18] A. Pronobis, O. Mozos, B. Caputo, SVM-based discriminative accumulation scheme for place recognition, in: Proceedings of the IEEE International Conference on Robotics and Automation, ICRA'08, Pasadena, CA, USA, May 2008.
[19] G. Griffin, A. Holub, P. Perona, Caltech-256 object category dataset, Technical Report 7694, Caltech, 2007, Available at: http://authors.library.caltech.edu/7694/.
[20] The PASCAL Visual Object Classes challenge, Available at: http://www.pascal-network.org/challenges/VOC/.
[21] The MIT-CSAIL database of objects and scenes, Available at: http://web.mit.edu/torralba/www/database.html.
[22] J. Ponce, T.L. Berg, M. Everingham, D. Forsyth, M. Hebert, S. Lazebnik, M. Marszałek, C. Schmid, C. Russell, A. Torralba, C. Williams, J. Zhang, A. Zisserman, Dataset issues in object recognition, in: Towards Category-Level Object Recognition, Springer, 2006, pp. 29–48.
[23] A. Howard, N. Roy, The Robotics Data Set Repository (Radish), 2003, Available at: http://radish.sourceforge.net/.
[24] E. Nebot, The Sydney Victoria Park dataset, Available at: http://www-personal.acfr.usyd.edu.au/nebot/dataset.htm.
[25] D.G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.
[26] O. Linde, T. Lindeberg, Object recognition using composed receptive field histograms of higher dimensionality, in: Proceedings of the 17th International Conference on Pattern Recognition, ICPR'04, Cambridge, UK, 2004.

[27] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, 2000.
[28] H. Tamimi, A. Zell, Vision based localization of mobile robots using kernel approaches, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS'04, Sendai, Japan, 2004.
[29] D. Filliat, A visual bag of words method for interactive qualitative localization and mapping, in: Proceedings of the IEEE International Conference on Robotics and Automation, ICRA'07, Roma, Italy, April 2007.
[30] J. Gaspar, N. Winters, J. Santos-Victor, Vision-based navigation and environmental representations with an omni-directional camera, IEEE Transactions on Robotics and Automation 16 (6) (2000).
[31] P. Blaer, P. Allen, Topological mobile robot localization using fast vision techniques, in: Proceedings of the IEEE International Conference on Robotics and Automation, ICRA'02, Washington, DC, USA, 2002.
[32] E. Menegatti, M. Zoccarato, E. Pagello, H. Ishiguro, Image-based monte-carlo localisation with omnidirectional images, Robotics and Autonomous Systems 48 (1) (2004).
[33] H. Andreasson, A. Treptow, T. Duckett, Localization for mobile robots using panoramic vision, local features and particle filter, in: Proceedings of the IEEE International Conference on Robotics and Automation, ICRA'05, Barcelona, Spain, 2005.
[34] A.C. Murillo, J.J. Guerrero, C. Sagues, SURF features for efficient robot localization with omnidirectional images, in: Proceedings of the IEEE International Conference on Robotics and Automation, ICRA'07, Roma, Italy, April 2007.
[35] C. Valgren, A.J. Lilienthal, Incremental spectral clustering and seasons: Appearance-based localization in outdoor environments, in: Proceedings of the International Conference on Robotics and Automation, ICRA'08, 2008.
[36] M. Mata, J.M. Armingol, A. de la Escalera, S.M.A., Using learned visual landmarks for intelligent topological navigation of mobile robots, in: Proceedings of the IEEE International Conference on Robotics and Automation, ICRA'03, 2003, Taipei, Taiwan.
[37] H. Bay, T. Tuytelaars, L. Van Gool, SURF: Speeded up robust features, in: Proceedings of the 9th European Conference on Computer Vision, ECCV'06, Graz, Austria, 2006.
[38] F. Fraundorfer, C. Engels, D. Nistér, Topological mapping, localization and navigation using image collections, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS'07, San Diego, CA, USA, October 2007.
[39] A. Torralba, P. Sinha, Recognizing indoor scenes, Technical Report 2001-015, AI Memo, 2001.
[40] A. Torralba, Contextual priming for object detection, International Journal of Computer Vision 53 (2) (2003).
[41] D.M. Bradley, R. Patel, N. Vandapel, S.M. Thayer, Real-time image-based topological localization in large outdoor environments, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS'05, Edmonton, Alberta, Canada, August 2005.
[42] C. Weiss, H. Tamimi, A. Masselli, A. Zell, A hybrid approach for vision-based outdoor robot localization using global and local image features, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS'07, San Diego, CA, USA, October 2007.
[43] M. Marszałek, I. Laptev, C. Schmid, Actions in context, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'09, 2009, Available at: http://www.irisa.fr/vista/actions/hollywood2/.
[44] The KTH-TIPS image database, Available at: http://www.nada.kth.se/cvap/databases/kth-tips/.
[45] A. Pronobis, B. Caputo, The KTH-INDECS database, Technical Report CVAP297, Kungliga Tekniska Hoegskolan, CVAP, September 2005, Available at http://cogvis.nada.kth.se/INDECS/.
[46] J. Luo, A. Pronobis, B. Caputo, P. Jensfelt, The KTH-IDOL2 database, Technical Report CVAP304, Kungliga Tekniska Högskolan, CVAP/CAS, October 2006, Available at http://cogvis.nada.kth.se/IDOL/.
[47] J. Folkesson, P. Jensfelt, H. Christensen, Vision SLAM in the measurement subspace, in: Proceedings of the IEEE International Conference on Robotics and Automation, ICRA'05, Barcelona, Spain, 2005, pp. 30–35.
[48] The Semantic Robot Vision challenge, URL: http://www.cs.cmu.edu/~srvc/.
[49] Image CLEF 2009 Robot Vision challenge, URL: http://imageclef.org/2009/robot/.
[50] K. Mikolajczyk, C. Schmid, Indexing based on scale invariant interest points, in: Proceedings of the 8th IEEE International Conference on Computer Vision, ICCV'01, Vancouver, Canada, 2001.
[51] K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'03, Madison, WI, USA, 2003.
[52] G. Dorkó, C. Schmid, Object class recognition using discriminative local features, 2005.
[53] M.E. Nilsback, B. Caputo, Cue integration through discriminative accumulation, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'04, Washington, DC, USA, 2004.
[54] B. Caputo, E. Hayman, P. Mallikarjuna, Class-specific material categorisation, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV'05, Beijing, China, 2005.
[55] A. Barla, F. Odone, A. Verri, Histogram intersection kernel for image classification, in: Proceedings of the International Conference on Image Processing, ICIP'03, Barcelona, Spain, 2003, pp. 513–516.
[56] J.C. Platt, Fast training of support vector machines using sequential minimal optimization, in: B. Schölkopf, C. Burges, A. Smola (Eds.), Advances in Kernel Methods: Support Vector Learning, MIT Press, Cambridge, MA, USA, 1999, pp. 185–208.
[57] O. Chapelle, P. Haffner, V. Vapnik, Support vector machines for histogram-based image classification, IEEE Transactions on Neural Networks 10 (5) (1999).
[58] C. Wallraven, B. Caputo, A. Graf, Recognition with local features: the kernel recipe, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV'03, Nice, France, 2003.
[59] S. Boughorbel, J.-P. Tarel, F. Fleuret, Non-mercer kernels for svm object recognition, in: Proceedings of the 15th British Machine Vision Conference, BMVC'04, London, England, September 2004.
[60] J.C. Platt, N. Cristianini, J. Shawe-Taylor, Large margin DAGs for multiclass classification, in: Advances in Neural Information Processing Systems, vol. 12, 2000, pp. 547–553.
[61] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, 2001, Software available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm.
[62] A. Pronobis, Indoor place recognition using support vector machines, Master's thesis, NADA/CVAP, Kungliga Tekniska Högskolan, Stockholm, Sweden, December 2005, Available at http://www.csc.kth.se/~pronobis/.
[63] K. Grauman, T. Darrell, The pyramid match kernel: Discriminative classification with sets of image features, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV'05, Beijing, China, October 2005.

A. Pronobis received his M.Sc. in Computer Science from the Silesian University of Technology in Gliwice, Poland in 2005. His thesis work was performed at the Royal Institute of Technology (KTH) in Stockholm, Sweden and focused on the use of kernel methods for visual place recognition. Since 2006, he pursues his Ph.D. at the Centre for Autonomous Systems at KTH where he works on multi-modal place classification and spatial knowledge representation in robotic systems. He is also involved in the CogX European project.

B. Caputo is a senior researcher at the IDIAP Research Institute since August 2006, where she works on robot learning, cognitive systems and visual categorization. She obtained her PhD in computer science from the Royal Institute of Technology, Stockholm, Sweden (2004), where she worked on probabilistic kernel methods for visual recognition. Her current research interests are in the field of visual recognition for robot systems, online learning for cognitive systems and audio-visual recognition of unexpected events. She is author/co-author of more than 40 journal and peer-reviewed conference papers.

P. Jensfelt received his M.Sc. in Engineering Physics in 1996 and Ph.D. in Automatic Control in 2001, from the Royal Institute of Technology, Stockholm, Sweden. Between 2002 and 2004 he worked as a project leader in two industrial projects. He is currently an assistant professor with the Centre for Autonomous System (CAS) and the principal investigator of the European project CogX at CAS. His research interests include mapping and localization and systems integration.

H.I. Christensen is the KUKA Chair of Robotics at the College of Computing, Georgia Institute of Technology. He is also the director of the Center for Robotics and Intelligent Machines (RIM@GT). Dr. Christensen does research on systems integration, human–robot interaction, mapping and robot vision. He has published more than 250 contributions across AI, robotics and vision.